Multi-view region-adaptive multi-temporal DMM and RGB action recognition

Al-Faris, Mahmoud ; Chiverton, John P.; Yang, Yanyan; Ndzi, David

Published in:
Pattern Analysis & Applications

DOI:
10.1007/s10044-020-00886-5

Published: 21/04/2020

Document Version
Publisher's PDF, also known as Version of record

Pattern Analysis and Applications
https://doi.org/10.1007/s10044-020-00886-5

THEORETICAL ADVANCES

Multi-view region-adaptive multi-temporal DMM and RGB action recognition

Mahmoud Al-Faris[1] · John P. Chiverton[1] · Yanyan Yang[2] · David Ndzi[3]

Received: 11 April 2019 / Accepted: 3 April 2020

Abstract
Human action recognition remains an important yet challenging task. This work proposes a novel action recognition system. It uses a novel multi-view region-adaptive multi-resolution-in-time depth motion map (MV-RAMDMM) formulation combined with appearance information. Multi-stream 3D convolutional neural networks (CNNs) are trained on the different views and time resolutions of the region-adaptive depth motion maps. Multiple views are synthesised to enhance the view invariance. The region-adaptive weights, based on localised motion, accentuate and differentiate parts of actions possessing faster motion. Dedicated 3D CNN streams for multi-time resolution appearance information are also included. These help to identify and differentiate between small object interactions. A pre-trained 3D-CNN is used here with fine-tuning for each stream along with multi-class support vector machines. Average score fusion is used on the output. The developed approach is capable of recognising both human action and human–object interaction. Three public-domain data-sets, namely MSR 3D Action, Northwestern UCLA multi-view actions and MSR 3D daily activity, are used to evaluate the proposed solution. The experimental results demonstrate the robustness of this approach compared with state-of-the-art algorithms.

Keywords Action recognition · DMM · 3D CNN · Region adaptive

1 Introduction

Action recognition is a key step in many amazing applications areas. Potential areas of interest are wide. They include automated security monitoring, [1]; social applications [2]; intelligent transportation [3]; smart hospitals [4]; and homes [5].

Action recognition methods can be based on a number of different sources of features such as space-time interest points [6], improved trajectories of features and fisher vectors [7, 8]. These techniques model motion in video data which are obviously an important source of information that can be used to help recognise actions. Instead of points of motion, less localised sources of motion can also be considered to model the motion of the body as a whole such as motion history images (MHIs) [9] and for the boundary as with motion boundary histograms (MBHs) [7]. Depth can also be incorporated with techniques such as depth motion maps (DMMs) [10].

These sources of, what might be considered handcrafted features are rich in information but not necessarily always able to capture all the relevant aspects of motion that might be needed to help a classifier to distinguish between different actions.

The introduction of deep learning techniques such as convolutional neural networks (CNNs) [11] presented significant advantages for many machine learning applications, not least computer vision including action recognition, see, e.g., [12]. Deep learning-based features extracted using, for example, CNNs have shown great performance over many traditional handcrafted features due to, in simple terms, their capability to learn the important aspects of actions from the huge amount of variation that can potentially occur in images and video sequences. This property has also enabled deep learning-based techniques to have improved invariance to, for example, pose, lighting and surrounding clutter [13]. It can also be seen that the inherent structure of CNN-based

✉ John P. Chiverton
john.chiverton@port.ac.uk

[1] School of Energy and Electronic Engineering, University of Portsmouth, Portsmouth PO1 3DJ, UK

[2] School of Computing, University of Portsmouth, Portsmouth PO1 3HE, UK

[3] School of Computing, Engineering and Physical Sciences, University of the West of Scotland, Paisley PA1 2BE, UK

Published online: 21 April 2020

techniques enables the preservation of the important relations in both the spatial and temporal dimensions [14]. As a part of the success of the deep learning-based methods, many variations in the architectures and approaches have been proposed.

1.1 Contributions

This work makes a number of novel contributions which are:

- Region-adaptive depth motion map (RA-DMM). Variable emphasis is placed on different regions in the motion maps with the aid of spatially localised estimates of motion using optical flow.
- A system that combines multi-synthesised views and multi-resolution motion information with multi-resolution appearance information (RGB) within a deep learning framework for action recognition. The appearance information is important for assisting with object interactions. Whilst the multi-resolutions assist with recognising the same actions performed with differing speeds. The synthesised views improve view invariance thus helping to further distinguish between actions.
- A hierarchical approach to action recognition in terms of recognising gross poses (i.e. standing, sitting, lying) and then specialised networks for the action recognition. This has the advantage of improving action recognition for the same action but in different poses. For instance, mobile phone usage whilst standing or sitting is more easily recognised given the gross pose information.

Section 2 gives an overview of related work, some disadvantages and some further information regarding the contributions that this work makes. Following that, the methodology in sect. 3 describes the approach proposed here in this paper. Section 4 presents the experiments and results. Finally, Section 5 presents the conclusions.

2 Related work

A number of techniques process single video frames as static CNN features [15, 16]. Others [15, 17, 18] have processed short video clips where video frames were employed as multi-channel inputs to 2D CNNs. A further development is the use of 3D CNNs where Ji et al. in 2013 [12] used 3D convolutions to incorporate both the spatial and temporal information of actions in video.

An extension to the conventional single-stream CNN model was proposed for the first time by Simonyan and Zisserman in 2014 [15] for action recognition. It used a two-stream approach to learn single-frame appearance information in combination with stacked optical flow of multiple frames which yielded improved performance.

More recently, deep learning techniques have increasingly been used to utilise temporal information for action recognition tasks. A unique architecture was proposed in [19] using a long-term recurrent CNN with both RGB and optical flow inputs.

Temporal periods over which temporal information is learned and recognised can be very short, e.g. 2 frames as in [20]. Incorporating more temporal information can help improve action recognition performance, as shown by, for example, [12, 16, 21], and multi-temporal resolution, as used by [14]. These methods utilised a range of different features but the advantage of the multi-temporal resolution approach is the ability to adapt to different actions carried out at different speeds.

A deeper 3D CNN network called C3D was built in [21], and the learned motion features used different massive public video data-sets. The features were shown to be compact and efficient as well as providing superior performance. The C3D model included eight convolution layers, five pooling layers, two fully connected layers.

In [22], a DMM-pyramid architecture was used to train both a traditional 2D CNN and 3D CNN to keep the partial temporal information of depth sequences for action recognition. The experiments achieved comparable results with state-of-the-art methods in terms of a number of different data-sets.

A CNN model obtained from ImageNet was used in [23]. It was used to learn from multi-view DMM features for action recognition where a video was projected onto different viewpoints within the 3D space. Different temporal scales were then used from the synthesised data to constitute a spatiotemporal pattern of an action. Finally, three fine-tuned models were employed independently on the resulting DMMs. However, a fixed number of temporal scales of DMM still made the spatiotemporal information limited to action sequences carried out over a limited range of time. This would also equally need more spatiotemporal information in order for it to be recognised. In addition, some actions included object interactions which might be very difficult to discern purely from raw depth data.

In [24], a 3D CNN structure was designed to capture spatiotemporal features for action recognition. A support vector machine (SVM) classifier was then used to classify actions based on the captured features. Experimental results showed some competitive results on the KTH action recognition data.

Similarly, a 3D CNN was proposed in [25] to automatically extract spatiotemporal features. Then, however, a recurrent neural network (RNN) was used to classify each sequence considering the learned features for each time step. The experiments on the KTH data-set demonstrated

impressive performance in comparison with state-of-the-art approaches. Another use of a 3D CNN was by Taylor et al. in 2010 [20] with a Restricted Boltzmann Machine to learn spatiotemporal features.

An efficient approach was proposed by Liu et al. in 2017 [26], which used a joint-pooled 3D deep convolutional descriptor applied to skeletal feature data on action recognition data. The experimental results demonstrated promising performance. Temporal information was exploited in [27], which used a deep long/short-term memory (LSTM) method on skeleton-based data sequences, which was then combined using a fusion-based approach with appearance information and employed for action recognition.

Deep learning-based action recognition was also presented in [28] using depth sequences and skeleton joint information combined. A 3D CNN structure was used to learn the spatiotemporal features from depth sequences, and then joint-vector features were computed for each sequence. Finally, the SVM classification results of the two types of features were fused for action recognition.

The 3D positional information in depth data can be further emphasised, as was done by [29] where multiple views were derived of the depth data. The authors applied it to dynamic depth images rather than incorporating it into a DMM formulation.

The formulation of the DMM has also been considered. For instance, in [30], the authors weighted the DMM based on a function that varied the amount of influence from more recent frames. In [31], the authors extended this to multiple functions. In another approach in [32], the authors combined wearable inertial sensor data with depth camera data to weight DMMs. This latter approach is interesting; however, it requires the individual to wear and provide an additional source of data. Furthermore, the motion information is not spatially localised.

All these different sources of features are useful but most of them do not consider the way the motion might be carried out over different ranges of time. For example, the number of frames used in the optical flow stacking ranged between 7 and 15 frames, such as 7, 10 and 15 frames as used in [12, 16, 33], respectively. This can be considered important in cross-actor and even for the same actor at different time points or similar. Appearance information is also not commonly used. Also, little attention is given to how different image regions that might be considered of higher relevance for different actions. Furthermore, they do not consider the effect of higher level information (e.g. pose) on the underlying learnt feature space.

At a lower level, it can also be considered preferable to obtain motion information from multiple contiguous frames in addition to the spatial information. Therefore, more suitable approaches are needed to capture extra temporal information as well as to keep the complexity of the model as low as possible. To this end, we propose a new hierarchical pose detection and action recognition system. The pre-trained C3D model is adapted here to learn multi-resolution features from both the spatial and temporal dimensions using different contiguous frames of RGB data. Furthermore, we propose an adaptive Multi-resolution depth motion map calculated across multiple views with important action information learned through the 3D CNN model to provide extra motion-based features that emphasise the significance of moved parts of an action. In addition, multi-resolution raw appearance information (i.e. RGB) is used to exploit various spatiotemporal features of the RGB scene which helps to capture more specific information that might otherwise be difficult to obtain from depth sequence information alone such as object interactions and finer image details. Our adaptive action recognition system is illustrated in Fig. 1.

Our automated system is developed and evaluated based on three well-known publicly available data-sets including the Microsoft Research (MSR) Action 3D data-set [34], the Northwestern UCLA Multiview Action 3D data-set [35] and the MSR daily activity 3D data-set [36]. The experimental results demonstrate the robustness of our approach compared with state-of-the-art algorithms.

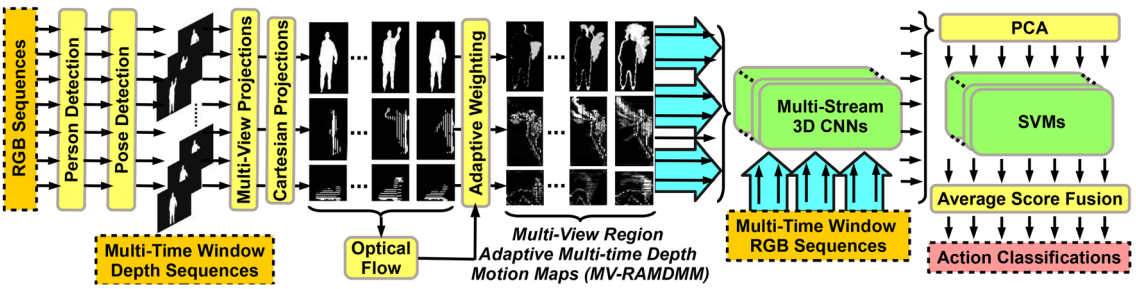

Fig. 1 Framework of our hierarchical region-adaptive multi-time resolution depth motion nap (RAMDMM) and multi-time resolution RGB action recognition system. Each pose, Cartesian projection, view and time window has a separate 3D CNN and SVM. The system is configured here to detect two poses, across seven views and three time resolutions in the three Cartesian planes. The RGB information is also detected across three time resolutions. This results in $2(3 \times 7 \times 3 + 3) = 132$ separate 3D CNNs and SVM classifiers

3 Methodology

Traditional depth motion maps (DMMs) are formulated on 2D planes by combining projected motion maps of an entire depth sequence. This does not consider the higher order temporal links between frames of depth sequences. A DMM can encapsulate a certain amount of the variation of a subject's motions during the performance of an activity. Unfortunately, difficulties can arise for activities that have the same type of movements but performed over different temporal periods. Our formulation therefore includes multiple time resolutions, referred to as Multi-resolution DMM (MDMM). Moreover, some actions or parts of actions are performed with different intensities. The differences in depth information captured at points of fast motion are accentuated using a region and motion-adaptive formulation producing a region-adaptive MDMM (RAMDMM). This adaptivity helps to further differentiate between actions, particularly with differences in depth due to positioning compared with actions with fast motion. Parameters used throughout this work are listed in Table 1.

3.1 Depth motion maps

The basic DMM (as used in, e.g., [10, 37, 38]) includes projecting each depth frame onto three orthogonal Cartesian planes. The motion energy from each projected view is then stacked. This can be through a specific interval or through the entire sequence to generate a depth motion map (DMM), Γ_v for each projection view,

$$\Gamma_v(t) = \sum_{t'=t}^{t+N-1} |m_v^{t'+1} - m_v^{t'}| \tag{1}$$

where $v \in \{xy, yz, xz\}$ indicates the Cartesian projection; m_v^t is the projected map of the depth information at time frame t under projection view v; and N is the number of frames that indicates the length of the interval. DMMs can be represented by combining the three generated DMMs Γ_v together where important information on body shape and motion is emphasised. Average score fusion is used here, to be discussed shortly in Sect. 3.3.

Table 1 Table of parameters and notation

Symbol	Description
$\Gamma_v(t)$	DMM, view v, time t
$v \in \{\text{xy,yz,xz}\}$	Views, front, side, top
N	Number of frames
m_v^t	Projected depth map, view v, time t
$\tau \in \{\lambda_1, \lambda_2, \lambda_3\}$	Window/temporal lengths
$\Gamma_{v,\tau}(t)$	DMM, view v, time t, temporal length τ
o_x, o_y	Depth map horizontal & vertical optical flow components
g_v^t, $g_{v,\alpha}^t$	Depth map magnitude optical flow
t'	Intermediate time value
$\mathbf{R}_x$, $\mathbf{R}_y$	Rotation matrices
α, β	Rotation angles
$\mathbf{p}$	Camera viewpoint
$\Gamma_{v,\tau}^{\text{of}}(t)$	DMM, view v, time t, temporal length τ, adaptively weighted
$\Gamma_{v,\tau,\alpha}^{\text{of}}(t)$	DMM, view v, time t, temporal length τ, adaptively weighted, view angle α
ζ_{ij}^{xyz}	Output of CNN layer i, feature map j, position x, y, z
b_{ij}	Feature map bias
w_{ijm}^{pqr}	Kernel value, layer i, feature map j from feature map m, position p, q, r
P_i, Q_i, R_i	ith layer kernel sizes
$\text{CNN}_{v,\tau,\alpha}^{\lambda}$, $\text{CNN}_{v,\tau,\alpha}$	3D CNN with λ frames
$\Gamma'_{v,\tau,\alpha}(t)$	Adaptive DMM for view v, window τ and angle α
$c_{v,\tau,\alpha}(t)$, $\text{SVM}_{v,\tau,\alpha}$	SVM result for view v, window τ and angle α
$c^{\text{dmm}}(t)$	Average score fusion over all motion SVM results for $v \in V$, $\tau \in \Lambda$ & $\alpha \in \mathscr{A}$
$c_r^{\text{rgb}}(t)$	RGB action classification (SVM) for r frames
$c^{\text{rgb}}(t)$	Average score fusion over all RGB SVM results
$r \in R$	RGB frame set
$c(t)$	Average score fusion between motion and appearance results

3.1.1 Multi-resolution-in-time depth motion maps

Mostly, a fixed number of frames have been used by other researchers or even the entire number of frames of an action sequence video to generate DMMs. But a length of an action is not known in advance. Hence, multi-resolution-in-time depth motion maps are needed to cover different temporal intervals and rates of an action.

To produce a Multi-resolution DMM (MDMM), the depth frames from a depth sequence are combined across three different ranges where each has a different time interval. This means that various values of temporal length τ are set to generate the MDMMs for the same action (depth sequence). As $\tau \in N^+$ in traditional DMMs, this can be improved by $\tau \in \{\lambda_1, \lambda_2, \lambda_3\}$ where $\lambda_i \in N^+$ are different temporal windows used to properly cover an action's motion regardless of whether it carries important information over a short or long duration. Each of these three durations produces a different DMM. The values of τ are selected to cover short, intermediate and long durations, where *long* would typically correspond to an entire depth sequence for the various video sequences considered here.

These MDMMs for each depth sequence can be calculated with:

$$\Gamma_{v,\tau}(t) = \sum_{t'=t}^{t+\tau-1} |m_v^{t'+1} - m_v^{t'}| \tag{2}$$

where $v \in \{xy, yz, xz\}$, $\tau = \lambda_i$ and, for example, $\tau \in \{5, 10, \text{All}\}$ (as used here) are the various lengths of depth sequence used to obtain an MDMM for each single frame.

3.1.2 Adaptive motion mapping

As already considered, different actions can be performed over different time periods. The MDMM is able to include motion information across a range of temporal windows. However, each action can also be performed at different speeds by different people and with movement in different locations in an image. Hence, an adaptive weighting approach based on the movement is applied to continuously weight the interest regions to adapt to any sudden change in an action.

To adapt various changes in an action, an adaptive weighting approach based on the magnitude of the optical flow motion vectors is employed to build a region-adaptive MDMM. Firstly, motion flow vectors are extracted using optical flow as explained in [39] on consecutive frames. Then, the motion magnitude for each single pixel is computed and normalised between two consecutive frames.

Optical flow is computed between two consecutive projected motion depth map frames, i.e. m_v^t and m_v^{t-1}. The result of the optical flow function is the motion flow vector $\boldsymbol{o}$ with vector elements $\boldsymbol{o}_x$ and $\boldsymbol{o}_y$ in the vertical and horizontal directions, respectively. The motion magnitude of the flow vectors of each pixel can be calculated using: $g = \boldsymbol{o}_x^2 + \boldsymbol{o}_y^2$. As the motion magnitude changes based on the type, speed and shape of an action movement, this can be utilised to improve the DMM calculation formula by including the motion magnitude in the DMM equation as a weighting function. This helps to add increased consideration for higher interest regions of a DMM template as well as providing low consideration for other regions. In addition, it can make the DMM template adapt to different movements in an action movement. The new RAMDMM can be formulated as follows:

$$\Gamma_{v,\tau}^{\text{of}}(t) = \sum_{t'=t}^{t+\tau-1} \left(|m_v^{t'+1} - m_v^{t'}| \times g_v^{t'+1} \right) \tag{3}$$

where $g_v^{t'+1}$ is the motion magnitude for view v at time point $t' + 1$. Figure 2 shows samples of DMM templates illustrating some differences between traditional DMMs and the region-adaptive DMM method.

3.2 Multiple views

The 3D characteristics of the depth sequences mean that it is possible to calculate different viewpoints of the same data. This can help to improve the model by making it view invariant. A virtual camera can be rotated with a specific value in 3D space, which can be seen to be equivalent to rotate the 3D points of the depth frames.

The virtual camera can be moved within the depicted space, for instance, from point $\mathbf{p}$ to $\mathbf{p}'$. The first step is to move from $\mathbf{p}$ to $\mathbf{p}_b$ with rotation angle α around Y-axis, then from $\mathbf{p}_b$ to $\mathbf{p}'$ with rotation angle β around X-axis. This is performed by the rotation matrices:

Fig. 2 Samples of traditional (top row) and adaptive weighted (bottom row) DMM templates (left to right): bend, tennis serve, forward kick and two hands wave

$$\mathbf{R}_x = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{bmatrix}, \tag{4}$$

and

$$\mathbf{R}_y = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{bmatrix}, \tag{5}$$

The right-handed coordinate system is used for the rotation where the original camera viewpoint is $\mathbf{p}$. Hence, the new coordinate of 3D point after rotation can be considered as follows:

$$\mathbf{p}' = \mathbf{R}_x \mathbf{R}_y \mathbf{p} \tag{6}$$

where $\mathbf{p}'$ is the new coordinate, and the corresponding depth value for the synthesised depth frames.

Our view projection method on depth sequences is similar to [29] except applied here to enable extraction of DMMs. Some results of multi-view projection are presented in Fig. 3 with different values of α rotation angles. It can be noticed that more discriminative information can be obtained by computing RA-DMM based on the synthesised depth frames.

Sequences of synthesised depth frames with different viewpoints can be synthesised from a series of these multi-view projections. This can contribute to better data augmentation for training processes in addition to better overall feature extraction.

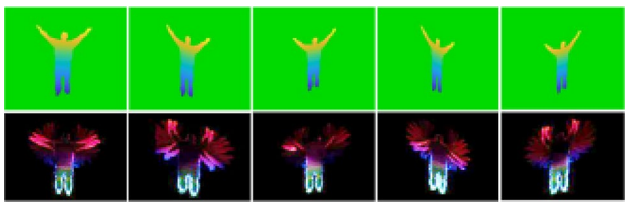

Fig. 3 Samples of original and synthesised depth frames (1*st* row) with RA-DMM (2*nd* row) after multiple viewpoints rotation when (left to right): $\alpha \in (0, 30, -30, 45, -45)$ respectively

In terms of the DMM formulation, multi-view extends the formulation with an additional dependency term, i.e.

$$\Gamma^{\text{of}}_{v,\tau,\alpha}(t) = \sum_{t'=t}^{t+\tau-1} \left(|m^{t'+1}_{v,\alpha} - m^{t'}_{v,\alpha}| \times g^{t'+1}_{v,\alpha} \right) \tag{7}$$

where $\alpha \in \mathscr{A}$ is from a sequence of angular values where $\mathscr{A} = (-45, ..., 45)$. Here, $\mathscr{A} = (-45, -30, -15, 0, 15, 30, 45)$ so that $|\mathscr{A}| = 7$.

3.3 Feature extraction, classification and fusion

An effective approach was presented for action recognition in [21] to learn spatiotemporal features using a 3D convolutional neural network which was also trained on a number of different large video data-sets. The training settings were kept the same as the original C3D model.

A 3D CNN is able to capture temporal information based on 3D convolution and pooling operations which are performed in the spatial and temporal dimensions.

The C3D network has eight convolution layers and five pooling layers that followed on from each other. Two fully connected layers and a softmax loss layer are used to recognise at the individual action label level. The numbers of kernels are 64, 128, 256, 256, 512, 512, 512, 512 for the convolution layers. The size of all kernels in the 3D CNN was set to $3 \times 3 \times 3$ with stride $1 \times 1 \times 1$. For the 3D pooling layers, the kernel sizes were set to $2 \times 2 \times 2$ with stride $2 \times 2 \times 2$ except for the first pooling layer which had a kernel size of $1 \times 2 \times 2$ and a stride of $1 \times 2 \times 2$ in order to preserve the temporal information at the early stages. The fully connected layers have 4096 output units each. The network structure is summarised in Fig. 4.

Conventionally, the value at position (x, y, z) on the jth feature map in the ith layer can be formulated as follows:

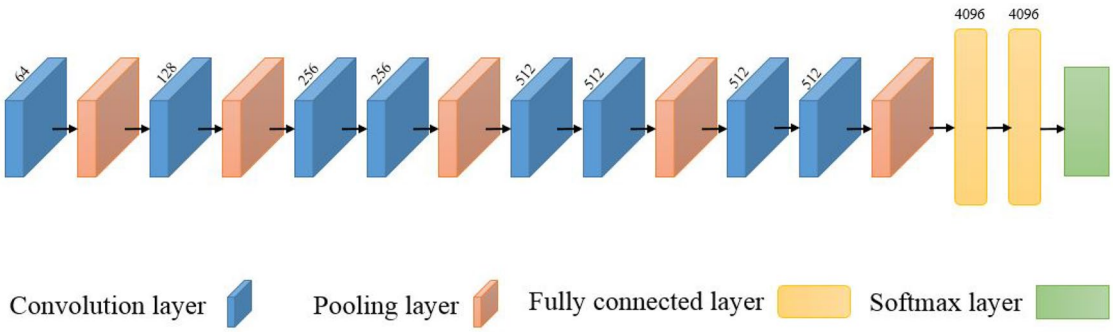

Fig. 4 A summary of the network structure used here. This structure was used for the MV-RAMDMM stream and again for the RGB stream

$$\zeta_{ij}^{xyz} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr}\zeta_{(i-1)m}^{(x+p)(y+q)(z+r)}\right), \quad (8)$$

where tanh(.) is the hyperbolic tangent function, b_{ij} is the bias for this feature map, m indexes over the set of feature maps in the $(i-1)$th layer connected to the current feature map, w_{ijm}^{pqr} is the value at the position (p, q, r) of the kernel connected to the m feature map in the previous layer. The kernel sizes R_i, P_i and Q_i are the temporal and spatial (height and width) dimensions, respectively.

Each value of the Cartesian projections v, time resolution τ and view α has a separate 3D CNN model that is trained based on a set of actions. The 2D output of each 3D CNN is then split into temporal feature vectors, and concatenation of the three orthogonal views is used to form a single feature vector. The dimensionality of each resulting feature vector is then reduced using Principal Component Analysis (PCA) determined from a covariance matrix of all the feature vectors. The projected feature vectors are then fed into different multi-class support vector machines (SVMs) [40] that are trained to recognise actions. The 3D CNN is trained to use a fixed number of input frames ($\lambda = 16$) for the depth information, i.e.

$$\text{CNN}_{v,\tau,\alpha}^{\lambda}(t) = \text{CNN}_{v,\tau,\alpha}\left(\Gamma'_{v,\tau,\alpha}(t), \Gamma'_{v,\tau,\alpha}(t-1), \ldots, \Gamma'_{v,\tau,\alpha}(t-\lambda)\right) \quad (9)$$

where $\Gamma'_{v,\tau,\alpha}(t)$ is a scaled and colour-mapped (jet) version of the multi-view region-adaptive multi-temporal resolution feature data $\Gamma^{\text{of}}_{v,\tau,\alpha}(t)$ for time t. The input frame size of the pre-trained C3D network is also fixed. A padding technique and interpolation are used here to resize frames to the required dimensions. Following the 3D CNN feature extraction process, feature concatenation and dimensionality reduction, the SVM classification is performed:

$$\boldsymbol{c}_{v,\tau,\alpha}(t) = \text{SVM}_{v,\tau,\alpha}\left(\text{CNN}_{v,\tau,\alpha}^{\lambda}(t)\right). \quad (10)$$

Classification vectors are then combined across all Cartesian planes, resolutions and views using average score fusion of the form:

$$\boldsymbol{c}^{\text{dmm}}(t) = \frac{1}{|V \times \Lambda \times \mathscr{A}|}\sum_{v}\sum_{\tau}\sum_{\alpha}\boldsymbol{c}_{v,\tau,\alpha}(t). \quad (11)$$

3.4 Multi-resolution spatiotemporal RGB information

Some types of actions and motions, especially those that interact with objects, can be perceived better with appearance information rather than, for example, depth due to the differences in the characteristics of the object in terms of appearances. In addition, it is somehow difficult to capture the DMM information of these objects, especially when the object's state is fixed or the size is relatively small.

Therefore, RGB data are utilised in this work as a source of the appearance information within our 3D CNN network model to capture discriminative spatiotemporal information of both subjects and interacting objects. Moreover, different temporal scales are used to cover different temporal ranges in the RGB scene, the same as for RAMDMM. This can help to mitigate against problems that might arise due to variations in the speed at which actions are performed that could result with different action performers. Three temporal scales are employed across three independently fine-tuned C3D models (in fixed mode for $\lambda = 16$ but then updated to use a variable number of inputs with $\lambda \in \{10, 25\}$), the outputs of which are fed into three independently trained multi-class support vector machines (SVMs). The outputs of the SVM classifiers are then combined together via average score fusion to form the multi-resolution RGB information:

$$\boldsymbol{c}^{\text{rgb}}(t) = \frac{1}{|R|}\sum_{r\in R}\boldsymbol{c}_r^{\text{rgb}}(t) \quad (12)$$

where $\boldsymbol{c}_r^{\text{rgb}}$ is the action classification vector for the RGB image frames taken across a time window of r frames.

An overall average score fusion is then used to derive the final classification vector, given by

$$\boldsymbol{c}(t) = \tfrac{1}{2}\left(\boldsymbol{c}^{\text{rgb}}(t) + \boldsymbol{c}^{\text{dmm}}(t)\right). \quad (13)$$

3.5 People detection and pose classification

The action recognition system can be made to perform well across a wide range of actions; however, this task can be further enhanced if the person performing an action can be localised in the image space. This helps remove extraneous background clutter and distractors. The performance of the system can also be further enhanced if the pose of the person can be detected prior to action recognition. It can be considered that this would help to provide the classification system with a better defined delineation between different actions performed in different poses. For instance, using a telephone whilst standing or sitting could produce a range of features that may not be that well connected in feature space or separated from other features from other actions.

Person detection is performed here using the Faster R-CNN [41] person detector based on the AlexNet [11] model as a network structure but transformed into a region proposal network (RPN), with the use of a ROI max pooling layer and classification layers.

A few samples from the RGB data of the utilised data-sets are used to create the ground-truth training data. After training, the created Faster R-CNN network is then used for person detection on the RGB data. This can help to eliminate the noise of the background environment in the action recognition process as can be seen in Section 4.

Pose detection is performed here using a specially adapted AlexNet pre-trained model [11] using transfer learning to classify the pose of an occupant out of three specified poses (sitting, standing and laying).

4 Experiments and results

Three public data-sets are used to evaluate the proposed method for action recognition: Northwestern UCLA Multiview Action 3D data-set [35]; Microsoft Research (MSR) Action 3D data-set [34]; and the MSR daily activity 3D data-set [36].

The overall steps and parameter values that are employed on the data-sets for feature extraction and action recognition are summarised as follows:

- Project the original depth sequence into different views with $\alpha \in (45, 30, 15, -15, -30, -45)$, which results in six synthesised views of the data and the original at $\alpha = 0$;
- Compute Cartesian projections of the seven views;
- Compute motion vectors' magnitude using optical flow algorithm over the original and synthesised sequences;
- Compute RA-DMMs for each sequence of original and synthesised sequences;
- Each action sequence is split into 16 frame subsequences in terms of RA-DMMs and 10-, 16-, 25-frame subsequences for the RGB information to train the 3D CNNs;
- Compute RA-DMM of each sequence of original and synthesised sequences using subsequence concatenation, dimensionality reduction and then average score fusion of three map templates;
- Multi-resolution RA-DMM computed using average score fusion for RA-DMM windows (5, 10, all);
- Multi-view RAMDMM computed using average score fusion of SVM classifiers for all RAMDMM across different views;
- Multi-resolution RGB (depth in MSR 3D action data-set) information computed using average score fusion of SVM classifiers for $\{10, 16, 25\}$ frames;
- Overall proposed system achieved with average score fusion between MV-RAMDMM and MR-RGB.

A summary of the parameter values can be seen in Table 2.

Table 2 Table of parameter values

Symbol	Values
$\tau \in \Lambda$	$\{5, 10, \text{All}\}$
$\alpha \in \mathscr{A}$	$(-45, -30, -15, 0, 15, 30, 45)$
λ	16 (C3D specific)
$r \in R$	For RGB where $R = \{10, 16, 25\}$

4.1 Northwestern UCLA data-Set

Northwestern UCLA Multiview action 3D data-set [35] has three Kinect cameras used to capture RGB, depth and human skeleton data simultaneously. This data-set includes ten different action categories: pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw and carry. Each action is performed by ten actors. In addition, this data-set consists of a variety of viewpoints.

We evaluate our proposed method with two different training and testing protocols for this data-set:

- Cross-subject training scenario: In this setting, we use the data of nine subjects as training data and leave the data of the remaining subject as test data. This is useful to show the performance of the recognition system across subjects. Furthermore, this is a standard criteria for comparison with the state of the art.
- Cross-view training scenario: As this data-set contains three view cameras, we use the data of two cameras as training data and leave the remaining camera as test data. This kind of setting is used to demonstrate the ability of the recognition system to perform with different views and to get another standard criteria to compare with the state of the art.

These settings give the opportunity to evaluate the proposed system with variations for different subjects and different views. The proposed method achieves an interesting set of results for the complete system demonstrating state-of-the-art performance as can be seen shortly. But first, let us examine the performance of the individual streams with individual inputs.

4.1.1 Multi-resolution-in-time appearance information

To start, the classification performance using multi-temporal resolution RGB data as an input to the 3D CNN model (C3D) is investigated together with the multi-class SVM classifier based on the aforementioned evaluation scenarios. Three temporal resolutions are used in terms of the RGB model including 10, 16, 25 windows. The trainable layers are adapted in the 3D CNN model when a nonconformant input

Table 3 Results of the fine-tuned C3D model with a multi-class SVM classifier for different time resolutions of the RGB data for the Northwestern UCLA data-set.

Settings	RGB_{10}	RGB_{16}	RGB_{25}	RGB_{fusion}
Cross-subject	67.44	78.12	88.23	91.51
Cross-view	56.71	61.79	70.32	72.20

Table 4 Results of the proposed model used RADMM and RAMDMM templates in terms of the Northwestern UCLA data-set

Settings	5	10	All	RAMDMM
Cross-subject	79.14	86.10	91.32	93.87
Cross-view	61.20	67.22	75.95	77.15

is used, i.e. $\lambda \in 10, 25$. An average score fusion is employed between the three SVMs to produce the multi-temporal resolution of the RGB data for action recognition. Table 3 includes the results of the different temporal resolutions for the RGB data in addition to the average score fusion result.

As we can see in Table 3, the fine-tuned C3D model achieves good performance in terms of cross-subject and cross-view classification schemes. The model already achieves relatively good recognition rates, particularly as the temporal window increases. It can be seen that C3D with multi-class SVM classification on RGB data alone with 25 temporal frames achieves the highest recognition performance of 88.23% and 70.32% in terms of cross-subject and view evaluation schemes, respectively. This reduces to 78.12%, 61.79% and 67.44%, 56.71% when 16 and ten temporal frames are used, respectively. Finally, the highest overall recognition performance is achieved when average score fusion is employed, combining the outputs of the three temporal results, again for RGB scene information only.

4.1.2 Multi-resolution-in-time region-adaptive depth motion maps

The region-adaptive DMM (RADMM) templates are calculated across the three temporal resolutions to form the multi-resolution DMM template, referred to as RAMDMM. These are used to learn discriminative features encapsulating depth, time and motion information. Results demonstrating the improvements achieved for the depth across multiple time windows are shown in Table 4. A similar trend as was seen for the appearance information can be observed for these results, i.e. a greater time window increases recognition performance, which is further improved by average score fusion for all time windows combined.

Table 5 Results of the proposed model used RAMDMM, MV-RAMDMM templates and average score fusion with MR-RGB in terms of the Northwestern UCLA data-set

Settings	RAMDMM	MV-RAMDMM	MV-(RAMDMM+RGB)
Cross-subject	93.87	96.30	97.15
Cross-view	77.15	84.52	86.20

Table 6 A comparison between the proposed method and state-of-the-art approaches in terms of Northwestern UCLA data-set

Paper	Cross-subject	Cross-view
Virtual view [42]	50.7	47.8
Hankelet [43]	54.2	45.2
MST-AOG [35]	81.6	73.3
Action Bank [44]	24.6	17.6
Poselet [45]	54.9	24.5
Denoised-LSTM [46]	–	79.6
tLDS [47]	93.0	74.6
MVDI [29]	–	84.2
kine-CNN [48]	–	75.6
R-NKTM [49]	–	78.1
Denoised-LSTM [50]	–	79.6
VE-LSTM [51]	–	87.2
E-TS-LSTM [52]	–	89.2
Ours	97.2	86.2

4.1.3 Combining RAMDMM-, multi-view- and appearance-based multiple sequences

The depth, time and motion information is then further combined across multiple synthesised views to produce MV-RAMDMM-based action recognition. At the end, an average score fusion is employed between the MR-RGB and MV-RAMDMM to utilise appearance, motion, shape and historical information based action recognition. Table 5 includes the results of the proposed method at different stages in the action recognition. The results in Table 5 appear to show that the different views of RAMDMM encapsulated within the MV-RAMDMM streams help to significantly improve the recognition rate for both the cross-subject and cross-view settings. In addition, an average score fusion between MV-RAMDMM and MR-RGB gives the opportunity to share a variety of important information for action recognition, improving the recognition accuracy in comparison with individual model classification reaching to 97.15 % and 86.20 % in terms of cross-view and cross-subject evaluation schemes, respectively. This can be compared to state-of-the-art approaches as seen in Table 6.

It can be seen in Table 6 that Virtual view [42] and Hankelet [43] methods are limited in their performance, which reflects the challenges of the Northwestern UCLA data-set (e.g. noise, cluttered backgrounds and various viewpoints). To mitigate these challenges, MST-AOG was proposed in [35] and achieved 81.60%. Our method achieves a significant improvement of 18% over MST-AOG and some comparable performance for the cross-view setting due to the big challenge in a cross-view setting. A confusion matrix of the proposed method is shown in Figure 5 using spatial and motion information in terms of the Northwestern UCLA multi-view action 3D data-set.

4.2 MSR 3D action data-set

The Microsoft Research (MSR) Action 3D data-set [34] is an action data-set consisting of depth sequences with 20 actions: high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw cross, draw tick, draw circle, hand clap, two hands wave, side-boxing, bend, forward kick, side kick, jogging, tennis serve, golf swing, pick up and throw. Each action is performed three times, each by ten subjects. A single point of view is used where the subjects were facing the camera whilst performing the actions. The data-set has been split into three groups based on complexity: AS1, AS2 and AS3 as used in many studies see, for example, [10, 34, 38, 53].

The action subsets are summarised in Table 7. All validation schemes make use of the three subsets.

Three evaluation schemes are considered in the literature (see, for example, [54]) in terms of the MSR action 3D data-set:

- One-third evaluation scheme: One-third of the instances are used as training samples and the reminder as testing samples. The one-third scheme splits the data-set using the first repetition of each action performed by each subject as training, and the rest for testing.
- Two-thirds evaluation scheme: Two-thirds of the instances are used as training samples and the remainder as testing samples. The two-thirds scheme splits the data-set into training samples using two repetitions of each action performed by each subject, and testing uses the rest of the data.
- Cross-subjects evaluation scheme: half of the subjects are used as training samples, and the other half are used as testing samples. Any half of the subjects can be used for testing, e.g. 2, 4, 6, 8 and 10, and the rest for training, i.e. 1, 3, 5, 7 and 9 (as used here).

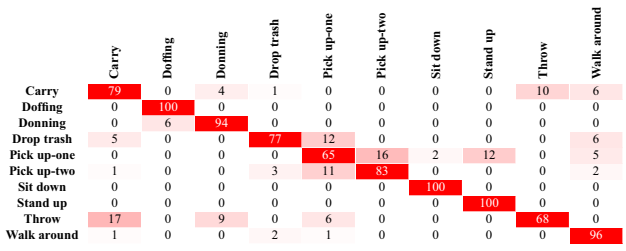

Fig. 5 Confusion matrices of the proposed method, using view-setting validation scheme in terms of Northwestern UCLA data-set

Table 7 Subsets of MSR action 3D data-set [34]

AS1	AS2	AS3
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw tick	Side kick
High throw	Draw cross	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two-hand wave	Tennis serve
Tennis serve	Side-boxing	Golf swing
Pick-up and throw	Forward kick	Pick-up and throw

Each subset has eight actions that can be used to evaluate the proposed method in terms of 1/3, 2/3 and cross-subject validation schemes. These can help to assess the performance of the proposed method against different training settings such as shortage of training samples, many training samples and variations between different subjects.

Similar to the experiments conducted above for the Northwestern UCLA 3D action data-set, a series of progressive sets of experiments are carried out.

4.2.1 Depth information

This data-set only has depth information (no appearance information). Therefore, instead of RGB-based appearance information, the depth frames are used. The pre-trained C3D network is individually implemented

Table 8 Performance of the C3D model based on multi-resolution depth information in terms of MSR 3D Action data-set

Subsets	Scheme	Depth_{10}	Depth_{16}	Depth_{25}	$\text{Depth}_{\text{fusion}}$
AS1	1/3	64.80	65.32	72.87	74.51
	2/3	75.30	76.71	77.14	80.10
	Cross	47.82	53.81	57.20	60.20
AS2	1/3	58.40	61.23	67.01	71.40
	2/3	61.72	68.18	74.89	76.72
	Cross	50.61	51.59	55.20	56.91
AS3	1/3	65.23	69.10	71.60	74.82
	2/3	69.17	78.43	80.94	81.10
	Cross	51.21	57.51	59.88	61.13

Table 9 Performance based on different lengths of RADMM, RAMDMM, average score fusion in terms of MSR 3D Action data-set

Subsets	Scheme	5	10	All	RAMDMM
AS1	1/3	72.10	79.34	95.32	96.53
	2/3	78.33	85.49	96.19	98.70
	Cross	62.95	63.87	85.50	88.31
AS2	1/3	74.42	77.13	94.11	95.90
	2/3	76.16	80.21	95.86	96.91
	Cross	57.39	62.04	80.27	83.89
AS3	1/3	76.89	81.56	95.87	97.20
	2/3	80.29	84.98	96.92	98.38
	Cross	63.70	66.87	87.16	90.42

based on depth data (instead of RGB) with various temporal frames 10, 16, 25 for the different MSR evaluation schemes. Then, an average score fusion is employed between the models to show the effect on the recognition rate. Table 8 includes the results of the C3D network implementation based on depth data alone.

Again, the recognition performance is improved with greater temporal windows and with different temporal dimensions combined by average score fusion, making the system more robust against speed variations. This demonstrates the utilisation of shape and temporal information from the depth sequences in the recognition process.

4.2.2 Multi-resolution-in-time region-adaptive depth motion maps

The performances of the multi-stream 3D CNNs, SVM classifiers and average score fusion across the different classifiers are now demonstrated for different lengths of the region-adaptive DMM (RA-DMM) templates on the MSR 3D action data-set. As before, these constitute the RA-DMM for multiple time resolutions to form the RAMDMM. Table 9 includes the results of the recognition model based on RADMM and RAMDMM information templates for different temporal windows.

The results in Table 9 includes the recognition performance for individual RADMM information along with multiple RADMM (RAMDMM). These results appear to show that learning actions' features based on RAMDMM is better than using either traditional DMM or individual length RADMMs. Moreover, it appears to show that sharing a variety of information available from the features by average score fusion between different models can improve the performance of the recognition system.

4.2.3 Combining RAMDMM-, multi-view- and depth-based multiple sequences

Table 10 shows the effects of the multi-view RAMDMM (MV-RAMDMM) templates and the effect of the multi-resolution spatiotemporal information on the recognition

Table 10 Performance of the recognition model based on RADMM, MV-RAMDMM, average fusion of MV-RAMDMM and MR-RGB C3D models in terms of MSR 3D Action data-set

Subsets	Scheme	RAMDMM	MV-RAMDMM	MV-(RAMDMM+depth)
AS1	1/3	96.53	97.90	99.21
	2/3	97.86	98.70	99.91
	Cross	88.31	91.28	97.95
AS2	1/3	95.90	97.40	99.08
	2/3	96.91	97.89	99.94
	Cross	83.89	89.11	95.89
AS3	1/3	97.20	98.33	99.89
	2/3	98.21	98.76	99.96
	Cross	90.42	94.80	95.77

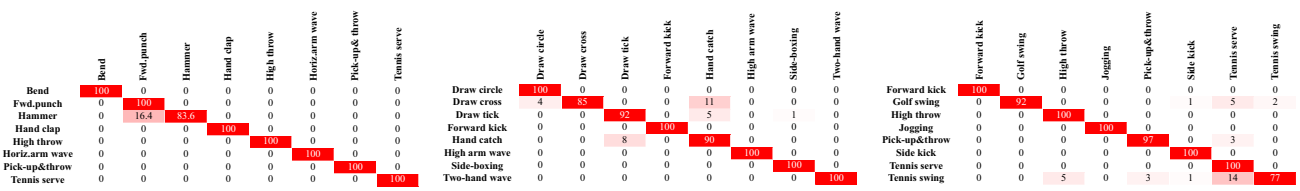

Fig. 6 Confusion matrices of the proposed method, using CS validation scheme in terms of AS1 (left), AS2 (middle) and AS3 (right) subset of MSR 3D action data-set

Table 11 Performance of the proposed method compared to the state-of-the-art approaches in terms of the MSR action 3D data-set [34]

Method	Accuracy %											
	1/3 Scheme				2/3 scheme				Cross subject scheme			
	AS1	AS2	AS3	Av.	AS1	AS2	AS3	Av.	AS1	AS2	AS3	Av.
Li et al. [34]	89.5	89.0	96.3	91.6	93.4	92.9	96.3	94.2	71.9	72.9	79.2	74.7
DMM-HOG [10]	97.3	92.2	98.0	95.8	98.7	94.7	98.7	97.4	96.2	84.1	94.6	91.6
Chen et al. [38]	97.3	96.1	98.7	97.4	98.6	98.7	100	99.1	96.2	83.2	92.0	90.5
HOJ3D [53]	98.5	96.7	93.5	96.2	98.6	97.2	94.9	97.2	88.0	85.5	63.6	79.0
Chaaraoui et al. [55]	–	–	–	–	–	–	–	–	91.6	90.8	97.3	93.2
DMM-HOG-KECA [56]	–	–	–	–	–	–	–	–	90.6	90.7	99.1	93.5
Vemulapalli et al. [57]	–	–	–	–	–	–	–	–	95.3	83.9	98.2	92.5
STOP [58]	98.2	94.8	97.4	96.8	99.1	97.0	98.7	98.3	91.7	72.2	98.6	87.5
DMM-LBP-FF [59]	96.7	100	99.3	98.7	100	100	100	100	98.1	92.0	94.6	94.9
DMM-LBP-DF[59]	98.0	97.4	99.3	98.2	100	100	100	100	99.1	92.9	92.0	94.7
tLDS [47]	–	–	–	–	–	–	–	–	96.81	89.14	98.83	94.85
CNN, SAE [60]	–	–	–	–	–	–	–	–	–	–	–	74.6
3D CNN, DHI [61]	–	-	–	–	–	–	–	–	–	–	–	92.8
VB-DMM [62]	98.0	97.4	99.3	98.2	98.6	100	100	99.5	99.1	92.3	98.2	96.5
DRN[63]	–	–	–	–	–	–	–	–	99.9	99.8	100	99.9
DMLAE[64]	–	–	–	–	–	–	–	–	–	–	–	84.0
Ours	99.2	99.1	99.8	99.3	99.9	99.9	99.8	99.9	97.9	95.8	95.7	96.5

accuracy of the system also combined with the depth sequences investigated in Section 4.2.1.

Figure 6 shows the confusion matrices of the recognition system using the proposed models under cross-subject evaluation schemes in terms of AS1, AS2 and AS3 subsets of MSR 3D action data-set.

Further, a comparison between the proposed method and the state-of-the-art approaches for human action recognition is presented in Table 11 in terms of the MSR Action 3D data-set under the aforementioned evaluation schemes.

It can be seen that our method outperforms the state-of-the-art approaches for the majority of cases and in others achieves at least comparable performance. Even though some of them are DMM-based methods such as [59] and [10], our method achieves greater recognition rate in the range of 1–6%. This appears to indicate that MV-RAM-DMM- and spatiotemporal information-based features can provide more powerful discrimination. Our approach utilises adaptive multiple hierarchical features that cover various periods of an action. In addition, the pre-trained recognition model uses a diverse range of layers which improves the chances to obtain the most accurate recognition performance.

4.3 MSR 3D Daily activity

The Microsoft Research (MSR) daily activity 3D data-set is among the most challenging data-sets because of a high level of intra-class variation, and many of the actions are based on object interaction. An action with object interaction is where the subject is interacting with an object when performing an action. This data-set has been captured by a Kinect sensor. It consists of depth and RGB sequences and includes 16 activities: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up and sit down. Performed by ten subjects, each subject performs an action twice in two different poses (standing and sitting).

Different evaluation schemes have been considered in the literature in terms of MSR daily activity 3D data-set. Here, similar to [23], a cross-subject validation is performed with subjects 1, 3, 5, 7, 9 for training and subjects 2, 4, 6, 8, 10 for testing. The person and pose detection steps are used to detect and localise a person within a frame, and pose detection is used to identify the pose, whether sitting or standing.

Table 12 Results of using multi-temporal resolution RGB data for the MSR 3D daily activity data-set

	RGB_{10}	RGB_{16}	RGB_{25}	RGB_{All}
Sit	53.73	57.50	64.48	65.90
Stand	51.20	56.81	61.11	63.79

Table 13 Results of RADMM, RAMDMM and MV-RMDMM with MR-RGB in terms of sitting pose of the MSR 3D daily activity data-set

5	10	All	RAMDMM	MV-RAM-DMM	MV-(RMDMM + RGB)
65.19	70.57	79.90	81.32	87.76	89.00

Table 14 Results of RADMM, RAMDMM and MV-RMDMM with MR-RGB in terms of standing pose of the MSR 3D daily activity data-set

5	10	All	RAMDMM	MV-RAM-DMM	MV-RMDMM+RGB
64.92	68.70	77.18	78.66	83.53	86.00

4.3.1 Multi-resolution-in-time appearance information

Firstly, multiple temporal resolutions (10, 16, 25) of RGB information are investigated separately with the fine-tuned C3D models. The outputs of these models are, as usual, classified using different SVMs. As before, the SVM outputs are combined using average score fusion. The results for this purely multi-temporal appearance-based recognition subsystem are shown in Table 12.

It can be seen in Table 12 that, as before, the robustness of the proposed model improves with an increase in the number of frames included in the system with the best combining the results from all temporal resolutions. This data-set is often considered to be much more complicated than others due to the two different scenarios for each single action, but the hierarchical strategy with the fine-tuned model is able to achieve comparable results based on RGB raw data. Moreover, a reasonable overall performance is also achieved that reaches 64.85% when an average recognition rate is employed.

4.3.2 Multi-resolution-in-time region-adaptive depth motion maps

As before, the RADMMs templates for three different temporal windows are computed and fed into fine-tuned C3D models, multi-class SVMs, the results of which constitute the RAMDMM for action recognition. Competitive results are achieved using these improved multiple temporal resolutions as can be seen in Table 13

For MV-RAMDMM, the performance reaches 89.00% and 86.00% within sitting and standing poses as presented in Table 14.

Further improvements can be seen by involving the multi-resolution spatiotemporal RGB information. Average score fusion improves the recognition of some objects' interaction actions and accomplishes 89% and 86% in terms of sitting and standing poses, respectively. The overall recognition rate of all data-sets can be calculated by taking the average of the two poses' recognition rates which reaches 87.5%. Figure 7 shows the confusion matrix of the hierarchical recognition system in terms of MSR 3D daily activity data-set.

A comparison between the proposed method and state-of-the-art approaches for action recognition is introduced in Table 15 in terms of MSR 3D daily activity data-set.

In Table 15, it can be seen that limited accuracy was previously achieved by LOP [36]- and DMM [10]-based approaches. Local HON4D was designed in [65] to tackle this kind of limitation and achieved a recognition rate of 80.00%. Actionlet Ensemble in [36] and SNV in [66] achieved a recognition rate that reaches 85.75% and 86.25%, respectively. These used a combination of depth and skeleton data. A recent method in [23] indicated the importance of DMM information and suggested the use of temporal depth motion maps and fine-tuned convolutional models. It achieved a relatively competitive result of 85.00%. Our method achieves comparable results with an improvement over some methods using our MV-RAMDMM and the spatiotemporal information of the C3D model. However, our method performed worse than the Range Sample [67] technique. This can be explained due to the noisy, complex and

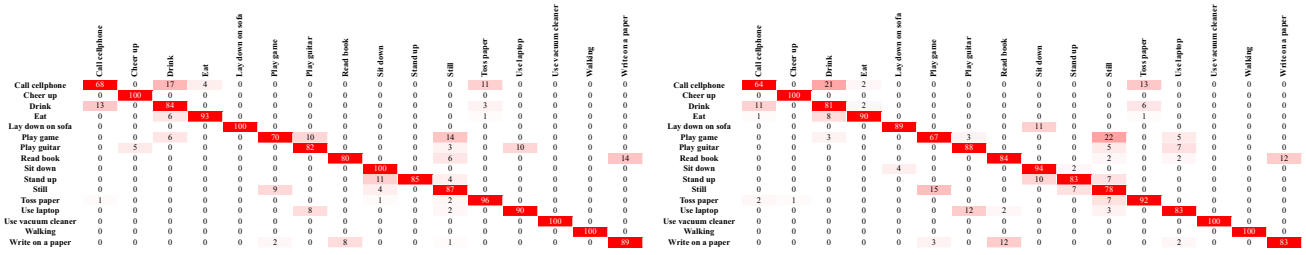

Fig. 7 Confusion matrices of the proposed method through standing pose (right) and sitting pose (left) in terms of daily activity 3D data-set

Table 15 Comparison of our method with the state-of-the-art approaches in terms of MSR daily activity 3D data-set [36]

Method	Accuracy %
LOP [36]	42.5
Depth motion maps [10]	43.1
Local HON4D [65]	80.0
Actionlet Ensemble [36]	85.8
SNV [66]	86.3
Range Sample [67]	95.6
DMM-CNN [23]	85.0
SNV [34]	86.3
DMMM [53]	81.9
DSTIP+DCSF [68]	83.6
WHDMM [69]	85.0
Actionlet Ensemble [70]	86.0
MDMMs [30]	89.0
MM2DCNN [71]	71.7
MMDT [71]	82.5
Deep Poselets [72]	84.4
DMLAE [64]	67.1
Ours	87.5

dynamic background of this data-set which can introduce significant noise in the RAMDMMs. Moreover, the Range Sample [67] method contained a technique that used skeleton data to eliminate the noise from the background. The confusion matrices in terms of MSR daily activity 3D data-set are shown in Fig. 7.

5 Conclusions

A novel feature representation technique for RGB-D data has been presented that enables multi-view and multi-temporal action recognition. A multi-view and multi-resolution region-adaptive depth motion maps (RA-DMMs) representation is proposed. The different views include the original and synthesised viewpoints to achieve view-invariant recognition. This work also makes use of temporal motion information more effectively. It integrates it into the depth sequences to help build in, by design, invariance to variations in an action's speed. An adaptive weighting approach is employed to help differentiate between the most important stages of an action. Appearance information in terms of multi-temporal RGB data is used to help retain a focus on the underlying appearance information that would otherwise be lost with depth data alone. This helps to provide sensitivity to interactions with small objects. Compact and discriminative spatiotemporal features are extracted using a series of fine-tuned 3D convolutional neural networks (3D CNNs). In addition, a pose estimation system is employed to achieve a hierarchical recognition structure. This helps the model to recognise the same action but with different positions. Multi-class support vector machines (SVMs) are used for action classification. Then, late score fusion technique is employed between different streams for the final decision.

The proposed method is robust enough to recognise human activities even with small differences in actions. This is in addition to achieving improved performance that is invariant to multiple viewpoints and providing excellent performance on actions that partly depend on human–object interactions. The system also remains invariant to a noisy environment and errors in the depth maps and temporal misalignments.

The proposed approach has been extensively validated on three benchmark data-sets: MSR 3D actions, Northwestern UCLA multi-view actions and MSR daily activities. The experimental results have demonstrated the great performance of the proposed method in comparison with state-of-the-art approaches.

Acknowledgements The authors would like to thank the anonymous reviewers for their valuable feedback. The first author (M.A.) acknowledges the support of the Higher Committee for Education Development in Iraq for funding his PhD study. This work also benefitted with credit provided by Google for using their Google Cloud.

References

1. Park S, Kim D (2018) Video surveillance system based on 3d action recognition. In: 2018 Tenth international conference on ubiquitous and future networks (ICUFN). IEEE, pp 868–870
2. Li Z, Tang J, Mei T (2018) Deep collaborative embedding for social image understanding. IEEE Trans Pattern Anal Mach Intell 41(9):2070–2083
3. Martin M, Roitberg A, Haurilet M, Horne M, Reiß S, Voit M, Stiefelhagen R (2019) Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In: Proceedings of the IEEE international conference on computer vision, pp 2801–2810
4. Khan I, Ahmed I, Ahmad M, Ullah K (2018) Towards a smart hospital: Automated non-invasive patient's discomfort detection in ward using overhead camera. In: 2018 9th IEEE annual ubiquitous computing, electronics and mobile communication conference (UEMCON). IEEE, pp 872–878

5. Shojaei-Hashemi A, Nasiopoulos P, Little JJ, Pourazad MT (2018) Video-based human fall detection in smart homes using deep learning. In: 2018 IEEE international symposium on circuits and systems (ISCAS). IEEE, pp 1–5
6. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Pattern recognition, 2004. ICPR 2004. Proceedings of the 17th international conference on, vol. 3. IEEE, pp 32–36
7. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, pp 3551–3558
8. Perronnin F, Sánchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification. In: European conference on computer vision. Springer, pp 143–156
9. Al-Faris M, Chiverton J, Yang L, Ndzi D (2017) Appearance and motion information based human activity recognition. In: Intelligent Signal Processing (ISP 2017), IET 3rd international conference on IET, pp 1–6
10. Yang X, Zhang C, Tian Y (2012) Recognizing actions using depth motion maps-based histograms of oriented gradients. In: Proceedings of the 20th ACM international conference on multimedia. ACM, pp 1057–1060
11. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25. Curran Associates Inc, New York, pp 1097–1105
12. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Machine Intell 35(1):221–231
13. Jing L, Ye Y, Yang X, Tian Y (2017) 3D convolutional neural network with multi-model framework for action recognition. In: 2017 IEEE international conference on image processing (ICIP). IEEE, pp 1837–1841
14. Varol G, Laptev I, Schmid C (2018) Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1510–1517
15. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems 27: annual conference on neural information processing systems 2014, Montreal, Quebec, Canada, 8–13 Dec 2014, pp 568–576
16. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1725–1732
17. Wang L, Xiong Y, Wang Z, Qiao Y (2015) Towards good practices for very deep two-stream convnets. arXiv preprint arXiv :1507.02159
18. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1933–1941
19. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
20. Taylor GW, Fergus R, LeCun Y, Bregler C (2010) Convolutional learning of spatio-temporal features. In: European conference on computer vision. Springer, Berlin, pp 140–153
21. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
22. Yang R, Yang R (2014) Dmm-pyramid based deep architectures for action recognition with depth cameras. In: Asian conference on computer vision. Springer, Berlin pp 37–49
23. Wang P, Li W, Gao Z, Zhang J, Tang C, Ogunbona PO (2016) Action recognition from depth maps using deep convolutional neural networks. IEEE Trans Hum Mach Syst 46(4):498–509
24. Latah M (2017) Human action recognition using support vector machines and 3D convolutional neural networks. Int J Adv Intell Inf 3(1):47–55
25. Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2011) Sequential deep learning for human action recognition. In: International workshop on human behavior understanding. Springer, pp 29–39
26. Liu H, Tu J, Liu M (2017) Two-stream 3D convolutional neural network for skeleton-based action recognition. arXiv, vol. arXiv :1705.08106
27. Das S, Koperski M, Bremond F, Francesca G (2018) A fusion of appearance based CNNs and temporal evolution of skeleton with LSTM for daily living action recognition. CoRR, vol. abs/1802.00421
28. Liu Z, Zhang C, Tian Y (2016) 3D-based deep convolutional neural network for action recognition with depth sequences. Image Vis Comput 55:93–100
29. Xiao Y, Chen J, Wang Y, Cao Z, Zhou JT, Bai X (2019) Action recognition for depth video using multi-view dynamic images. Inf Sci 480:287–304
30. Chen C, Liu M, Liu H, Zhang B, Han J, Kehtarnavaz N (2017) Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition. IEEE Access 5:22590–22604
31. Al-Faris M, Chiverton J, Yang Y, Ndzi D (2019) Deep learning of fuzzy weighted multi-resolution depth motion maps with spatial feature fusion for action recognition. J Imaging 5(10):82
32. Dawar N, Kehtarnavaz N (2018) Action detection and recognition in continuous action streams by deep learning-based sensor fusion. IEEE Sens 18(23):9660–9668
33. Park E, Han X, Berg T, Berg A (2016) Combining multiple sources of knowledge in deep CNNs for action recognition. In: Applications of computer vision (WACV), IEEE Winter conference, pp 1–8
34. Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3D points. In: 2010 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 9–14
35. Wang J, Nie X, Xia Y, Wu Y, Zhu S-C (2014) Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2649–2656
36. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1290–1297
37. Chen C, Jafari R, Kehtarnavaz N (2016) A real-time human action recognition system using depth and inertial sensor fusion. IEEE Sens J 16(3):773–781
38. Chen C, Liu K, Kehtarnavaz N (2016) Real-time human action recognition based on depth motion maps. J Real-time Image Process 12(1):155–163
39. Liu C, Freeman WT, Adelson EH, Weiss Y (2008) Human-assisted motion annotation. In: Computer vision pattern recognition, 2008. CVPR 2008. IEEE Conference IEEE pp 1–8
40. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) LIBLINEAR: a library for large linear classification. J Mach Learn Res 9:1871–1874
41. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 6:1137–1149

42. Li R, Zickler T (2012) Discriminative virtual views for cross-view action recognition. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2855–2862
43. Li B, Camps OI, Sznaier M (2012) Cross-view activity recognition using hankelets. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1362–1369
44. Sadanand S, Corso JJ (2012) Action bank: A high-level representation of activity in video. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1234–1241
45. Maji S, Bourdev L, Malik J (2011) Action recognition from a distributed representation of pose and appearance. In: 2011 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 3177–3184
46. Demisse GG, Papadopoulos K, Aouada D, Ottersten B (2018) Pose encoding for robust skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 188–194
47. Ding W, Liu K, Belyaev E, Cheng F (2018) Tensor-based linear dynamical systems for action recognition from 3D skeletons. Pattern Recognit 77:75–86
48. Wang J, Liu Y (2018) Kinematics features for 3D action recognition using two-stream CNN. In: 2018 13th world congress on intelligent control and automation (WCICA). IEEE, pp 1731–1736
49. Rahmani H, Mian A, Shah M (2017) Learning a deep model for human action recognition from novel viewpoints. IEEE Trans Pattern Anal Mach Intell 40(3):667–681
50. Demisse G, Papadopoulos K, Aouada D, Ottersten B (2018) Pose encoding for robust skeleton-based action recognition. In: Proceedings of IEEE conference computer visual pattern recognition workshop, pp 188–194
51. Baptista R, Ghorbel E, Papadopoulos K, Demisse GG, Aouada D, Ottersten B (2019) View-invariant action recognition from RGB data via 3D pose estimation. In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2542–2546
52. Lee I, Kim D, Kang S, Lee S (2017) Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. In: Proceedings of the IEEE international conference on computer vision, pp 1012–1020
53. Xia L, Chen C-C, Aggarwal J (2012) View invariant human action recognition using histograms of 3D joints. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 20–27
54. Padilla-López JR, Chaaraoui AA, Flórez-Revuelta F (2014) A discussion on the validation tests employed to compare human action recognition methods using the MSR action 3D dataset. arXiv preprint arXiv:1407.7390
55. Chaaraoui AA, Padilla-López JR, Climent-Pérez P, Flórez-Revuelta F (2014) Evolutionary joint selection to improve human action recognition with RGB-D devices. Expert syst Appl 41(3):786–794
56. El Madany NED, He Y, Guan L (2015) Human action recognition using temporal hierarchical pyramid of depth motion map and KECA. In: 2015 IEEE 17th international workshop on multimedia signal processing (MMSP). IEEE, pp. 1–6
57. Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3D skeletons as points in a lie group. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 588–595
58. Vieira AW, Nascimento ER, Oliveira GL, Liu Z, Campos MF (2014) On the improvement of human action recognition from depth map sequences using space-time occupancy patterns. Pattern Recogn Lett 36:221–227
59. Chen C, Jafari R, Kehtarnavaz N (2015) Action recognition from depth sequences using depth motion maps-based local binary patterns. In: 2015 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1092–1099
60. Tomas A, Biswas K (2017) Human activity recognition using combined deep architectures. In: 2017 IEEE 2nd international conference on signal and image processing (ICSIP). IEEE, pp 41–45
61. Keçeli AS, Kaya A, Can AB (2018) Combining 2D and 3D deep models for action recognition with depth information. Signal Image Video Process 12(6):1197–1205
62. Jin K, Jiang M, Kong J, Huo H, Wang X (2017) Action recognition using vague division DMMS. J Eng 4:77–84
63. Pham H-H, Khoudour L, Crouzil A, Zegers P, Velastin SA (2018) Exploiting deep residual networks for human action recognition from skeletal data. Comput Vision Image Understand 170:51–66
64. Yin X, Chen Q (2016) Deep metric learning autoencoder for nonlinear temporal alignment of human motion. In: 2016 IEEE international conference on robotics and automation (ICRA). IEEE pp 2160–2166
65. Oreifej O, Liu Z (2013) Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 716–723
66. Yang X, Tian Y (2017) Super normal vector for human activity recognition with depth cameras. IEEE Trans Pattern Anal Mach Intell 39(5):1028–1039
67. Lu C, Jia J, Tang C-K (2014) Range-sample depth feature for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition pp 772–779
68. Xia L, Aggarwal J (2013) Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: 2013 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2834–2841
69. Wang P, Li W, Gao Z, Zhang J, Tang C, Ogunbona P (2016) Action recognition from depth maps using deep convolutional neural networks. IEEE Trans Human Mach Syst 46:498–509
70. Wang J, Liu Z, Wu Y, Yuan J (2013) Learning actionlet ensemble for 3D human action recognition. IEEE Trans Pattern Anal Mach Intell 36(5):914–927
71. Asadi-Aghbolaghi M, Bertiche H, Roig V, Kasaei S, Escalera S (2017) Action recognition from RGB-D data: comparison and fusion of spatiotemporal handcrafted features and deep strategies. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3179–3188
72. Mavroudi E, Tao L, Vidal R (2017) Deep moving poselets for video based action recognition. In: 2017 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 111–120